Ground Vehicle Classification using Hidden Markov Models

William J.J. Roberts Howard W. Sabrin Yariv Ephraim

Atlantic Coast Technologies, Inc.
11499 Columbia Pike
Silver Spring, MD, 20904
robertsb,sabrinh@acoast.com

Dept. of ECE
George Mason University
Fairfax, VA, 22030
yephraim@gmu.edu

October 4, 2001

Abstract

Ground vehicle classification is performed using hidden Markov modelling of cepstral coefficients. The hidden Markov model (HMM) is used to represent audio signals. These signals are obtained as the vehicles travel past audio sensor arrays. Well known HMM training algorithms are applied to train models from training data. The trained models are used in two classification rules: the MAP rule, and a list-based rule due to Forney. Under some general assumptions, these approaches can be regarded as optimal. Using recordings from the ACIDS database, over 96% recognition rate on single vehicle classification is achieved. Multi-vehicle recordings from this database were simulated and good classification results obtained.

1 Introduction

The acoustic emissions produced by a ground vehicle may be used to classify that vehicle into its type. This capability may find application in the monitoring of militarily sensitive regions. If the system detects a military vehicle, a human operator can be alerted and directed munitions may be deployed. Such a system may eventually be able to address some of the requirements currently fulfilled by land-mines.

Ground vehicle classification is an example of a hypothesis testing problem. An optimal decision rule, in the sense of minimizing the probability of classification error, is given by the maximum a posteriori decision (MAP) rule. This rule is applied widely in may classification problems when a single decision is required. This is the case when signals from single vehicles are being classified. In real world situations, vehicles may travel in convoys consisting of multiple vehicles of multiple types. In such cases, the Forney decision rule [1] may be better suited. In this rule, all vehicles whose corresponding discrimination functions are greater than a threshold are placed on a list. The vehicles on this list constitute the guesses for the vehicles appearing in the test signal. Forney shows this rule to be optimal in Neyman-Pearson like sense.

Whichever decision rule is used, the probability density functions (pdf's) of the audio signals are required to implement the rule. These pdf's are not explicitly available and must be estimated from training data. The estimated pdf's can then be used in the decision rules as if they were the true pdf's. This technique is referred to as the "plug-in" technique and its optimality is discussed in [2, 3].

In this work, the pdf's of the audio signals are assumed to be hidden Markov models (HMM's). The HMM consists of a sequence of states that are visited in a Markovian manner. Each state of the HMM may be regarded as representing a particular sound from the

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
04 Oct 2001	N/A	-

Title and Subtitle Ground Vehicle Classification Using Hidden Markov Models	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Atlantic Coast Technologies, Inc. 11499 Columbia Pike Silver Spring, MD 20904	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) Department of the Army, CECOM RDEC Night Vision & Electronic Sensors Directorate AMSEL-RD-NV-D 10221 Burbeck Road Ft. Belvoir, VA 22060-5806	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
See also ADM201471, Papers from the Meeting of the MSS Specialty Group on Battlefield Acoustic and Seismic Sensing, Magnetic and Electric Field Sensors (2001) Held in Applied Physics Lab, Johns Hopkins Univ, Laurel, MD on 24-26 Oct 2001. Volume 2 (Also includes 1999 and 2000 Meetings), The original document contains color images.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
10

vehicle. Similarly to the use of HMM's in speech recognition, we use HMM's with state pdf's that are Gaussian with non-zero means and diagonal covariances.

In this, as in many other applications, it is not the time domain signal that is modelled as an HMM. Rather, as in speech recognition, the HMM models a feature vector obtained from the time domain signal. In this work, the elements of this feature vector are cepstral coefficients. Cepstral coefficients are calculated from the inverse Fourier transform of the logarithm of a spectral estimate of the signal. They result in high performance while using a low dimensional representation of the signal.

The remainder of the paper is organized as follows. In Section 2 we discuss the properties of cepstral coefficients. In Section 3 we give some salient details of the HMM. In Section 4 we provide details of the decision rules. In Section 5 we discuss the implementation and results. In Section 6 we give some comments.

2 Cepstral coefficients

The cepstrum is an example of a homomorphic [4] signal processing technique. It is defined as the inverse discrete Fourier transform of the log of the power spectral density of the signal. The cepstral sequence $c(n)$ corresponding to the power spectral density $S(\omega)$ is given by

$$c(n) = \int_{-\pi}^{\pi} \log S(\omega) e^{j\omega n} \frac{d\omega}{2\pi}. \tag{1}$$

Cepstral coefficients have a number of important properties that make them useful for classification applications.

Spectral Change Rate Information Low order cepstral coefficients capture information about the slowly varying properties of the spectrum. This is analogous to low order spectral coefficients capturing information about the slowly varying waveform. The slowly varying components of the spectrum are often referred to as the spectral envelope.

Gain Invariance Multiplication of the underlying signal by a constant gain will affect only the $c(0)$ term. The feature vector can thus be made invariant to changes in gain by exclusion of this term. Gain invariance is a highly desirable property in applications where classification needs to be performed in the face of arbitrary changes in gain of the underlying signal. This occurs in ground vehicle classification as the gain of the signal depends on the distance between the vehicle and the audio senors, and is therefor highly varying.

Known Statistical Properties In [5], cepstral coefficients obtained from a smoothed spectral estimate and auto-regressive spectral estimates were considered. In [6] more explicit non-asymptotic results were obtained for Gaussian signals and periodogram spectral estimate. In both cases it was shown that the asymptotic covariance of cepstral coefficients is a diagonal fixed signal independent matrix These properties justify the generally made choice of diagonal matrices for the covariances of the Gaussian mixtures in the HMM.

De-convolution Ability Cepstral coefficients have a ability to de-convolve signals, and thus potentially reduce unwanted channel effects. This property is critical in ground vehicle identification since it can be used to compensate for different audio sensors of for signal reverberation. Assuming a signal u results from passing an excitation w through a linear filter with impulse response g, then

$$u = w \otimes g \tag{2}$$

where $\otimes$ denotes convolution. In the frequency domain this is represented as

$$U(\omega) = W(\omega)G(\omega) \tag{3}$$

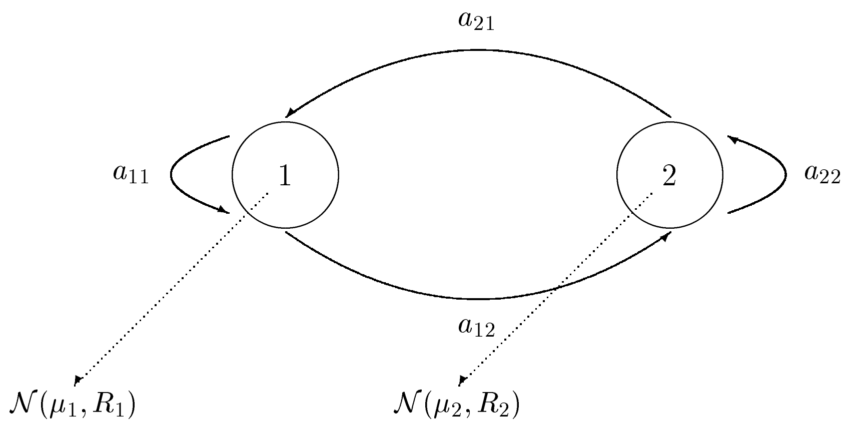

Figure 1: A two state Gaussian hidden Markov model

where $U(\omega), W(\omega)$, and $G(\omega)$ are the Fourier transforms of u, w, and g respectively. Taking the logarithm of both sides

$$\log U(\omega) = \log W(\omega) + \log G(\omega). \tag{4}$$

Hence in the log frequency domain, the component due to the excitation and due to the filter are additive and they may be separated using cepstral mean subtraction [7].

Estimation of the cepstral coefficients requires an estimate of the spectrum $S(\omega)$. A large amount of advice on the spectral estimation is available to guide the practitioner, see for example [8, 9]. Here we use Grenander and Rosenblatt's "window method" of spectral estimation [8, 6]. This method results in a consistent spectral estimate. The spectral estimate is obtained from the Fourier transform of a windowed auto-correlation estimate. The window used should result in a spectral estimate that is non-negative. Not all windows satisfy this property, some of them that do are listed in [8, 5.2.3]. Once a spectral estimate has been obtained, the cepstral coefficients are obtained from Fourier transform of the logarithm of the estimate. The method is more fully analyzed in [6].

3 Hidden Markov models

An HMM consists of a set of states each with an associated probability density function. At any given time instant, an output process is generated from a particular state. The identity of the state is not known. Intuitively, if we regard the modelled signal as consisting of a number of distinct sounds, then each state represents a statistical description of each of these sounds. With the passage of time, Markovian state transitions occur resulting in a sequence of states. These transitions are Markovian. As mentioned earlier, here we use state pdf's that are Gaussian with diagonal covariance matrices t0 represent the sequence of feature vector of cepstral coefficients.

Figure 1 shows a two state HMM with Gaussian output pdf's. The process begins in state 1 with probability π_1 or in state 2 with probability $\pi_2 = 1 - \pi_1$. At each time increment, the process either stays in the same state that it was in or it changes state according to a transition probability. The state transition probability from the state at time $t-1$, denoted by s_{t-1}, to the state at time t, s_t, is denoted by $a_{s_{t-1}s_t}$. Once in the state s_t, a K-variate Gaussian process is generated with state dependent mean and covariance $(\mu_{s_t},\ R_{s_t})$. Thus the parameter of the HMM is given by $\lambda = (\pi, a, \mu, R)$, where $\pi = \{\pi_\beta\}$, $a = \{a_{\alpha\beta}\}$, $\mu = \{\mu_\beta\}$, and $R = \{R_\beta\}$ for $\alpha, \beta = 1, \ldots, M$ where M is the number of HMM states.

We now present the standard assumptions of the HMM. For notational convenience, we suppress the conditioning of the parameter of the HMM on the particular hypothesis, as all hypotheses may be treated equally. Let $y = \{y_t, t = 1, \ldots, T\}, y_t \in \Re^K$, be a sequence of vectors generated by an HMM. Let $s = \{s_t, t = 1, \ldots, T\}, s_t \in \{1, \ldots, M\}$, be the sequence of states that generated y. We can express the acoustic model $p(y|\lambda)$ as

$$p(y|\lambda) = \sum_{s \in \mathcal{S}} p(y|s, \lambda) p(s|\lambda) \tag{5}$$

where $\mathcal{S}$ is the set of all possible state sequences, $p(y|s, \lambda)$ is the pdf of y given the state sequence s, and $p(s|\lambda)$ is the pmf of the state sequence. Using the assumption that the state transitions are first order Markovian we have

$$p(s|\lambda) = \prod_{t=1}^{T} a_{s_t s_{t-1}} \tag{6}$$

where $a_{s_t s_{t-1}}$ is the transition probability from state s_{t-1} to state s_t. The observation vectors $\{y_t\}$ are assumed to be independent of each other given the state sequence $\{s_t\}$. Thus

$$p(y|s, \lambda) = \prod_{t=1}^{T} p(y_t|s_t, \lambda) \tag{7}$$

and hence

$$p(y|\lambda) = \sum_{s \in \mathcal{S}} \prod_{t=1}^{T} a_{s_t s_{t-1}} p(y_t|s_t, \lambda). \tag{8}$$

The parameter of the HMM is estimated from training data. A computationally efficient algorithm, due to Baum *et al* [10, 11], is available for the ML estimate of λ. Baum's algorithm is iterative and is an example of, what was later known as, the expectation-maximization (EM) approach [12]. Other training approaches are possible, *e.g* MMI, MDI, or minimizing the empirical error rate, but their implementation is significantly more complicated than the ML approach.

The estimated pdf's are subsequently used in the decision rules as if they were the true pdf's. Optimality of this approach is discussed in [3].

4 Decision Rules

Assuming that all pdf's are explicitly known, we consider two decision rules. The first is the maximum a posteriori (MAP) rule [13]. This test is optimal in the minimum probability of error sense. The second is a list based rule due to Forney [1]. This rule is optimal in a generalized Neyman-Pearson sense. Forney's rule is used for the multi-vehicle problem as it allows a list of vehicles to be produced for a given acoustic signal.

4.1 MAP Rule

The MAP rule is frequently applied in classification problems. Given a signal y to be classified, the MAP rule chooses the ith hypothesis H_i by

$$\max_{\{H_i\}} p(y|H_i) p(H_i) \tag{9}$$

where $p(y|H_i)$ is the pdf of the signal from hypothesis H_i and $p(H_i)$ is the a priori probability of that hypothesis. The MAP rule exhaustively partitions the decision space (See Figure 2.a). into disjoint regions. This rule in suited for a situation where y resulted from a single vehicle.

4.2 List-Based Rule

Forney's rule was originally obtained for finite-alphabet problems, but it is equally applicable to the continuous-alphabet case. Given a signal to be classified, those hypotheses with a discrimination function greater than a threshold η are placed on a list. If no discrimination function is greater than the threshold, no decision is made and the signal is not classified. Using a generalized Neyman-Pearson lemma, Forney derived the decision rule that maximized the probability that the correct hypothesis is on the list for a given number of hypotheses erroneously on the list. Let H_j be the hypothesis that the jth vehicle type generated the test signal. Let $\Lambda_j(y), j = 1, \ldots, J$ be the discrimination functions and $\Omega_j = \{y : \Lambda_j(y) > \eta)\}$ be the jth decision region, where η represents a threshold. These $\{\Omega_j\}$ are not assumed to be disjoint, see Figure 2, so more than one hypothesis can be placed on the list. Let $\bar{N}$ equal the average

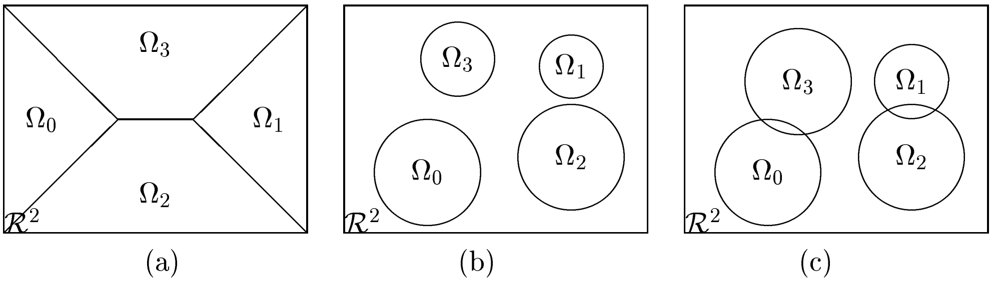

Figure 2: Decision regions in $\mathcal{R}^2$: (a) MAP test, (b) disjoint decision regions not covering the entire space, (c) overlapping regions not covering the entire space.

number of incorrect entries on the list and P_D equal the probability that the correct hypothesis appears on the list. We have

$$\begin{aligned}
\bar{N} &= \sum_{j=1}^{J} \text{Pr } (H_j \text{ erroneously on the list}) \\
&= \sum_{j} \sum_{j' \neq j} \text{Pr } (y \in \Omega_j \text{ and } H_{j'} \text{ is true}) \\
&= \sum_{j} \sum_{j' \neq j} p(H_{j'}) \int_{y \in \Omega_j} p(y|H_{j'}) dy \\
&\leq J - 1
\end{aligned} \tag{10}$$

and

$$\begin{aligned}
P_D &= \sum_{j=1}^{J} \text{Pr } (H_j \text{ correctly on the list}) \\
&= \sum_{j=1}^{J} \text{Pr } (y \in \Omega_j \text{ and } H_j \text{ is true}) \\
&= \sum_{j=1}^{J} p(H_j) \int_{y \in \Omega_j} p(y|H_j) dy \\
&\leq 1
\end{aligned} \tag{11}$$

In [1] a generalized Neyman-Person lemma is proved to find the decision rule that maximizes P_D for a given bound δ on $\bar{N}$ *i.e.*

$$\max_{\{\Omega_j\}} P_D \quad \text{s.t.} \quad \bar{N} \leq \delta \tag{12}$$

This yields the following optimal discrimination function

$$\Lambda_j(y) = \frac{1}{T} \log \frac{p(H_j)p(y|H_j)}{\sum_{j' \neq j} p(H_{j'})p(y|H_{j'})} \qquad j = 1, \ldots, J \tag{13}$$

and H_j is placed on the list if the observation y satisfies $\Lambda_j(y) > \eta$.

The loci of $\bar{N}$ and P_D for various thresholds produces a plot akin to the receiver operating characteristic (ROC) curve [13]. Generally, it is the empirical values of these quantities obtained from unseen test utterances that are useful as performance metrics. In Appendix A we show how to calculated empirical values given test utterances.

5 Experimental Results

In this section we describe the implementation and testing of a HMM based vehicle classification system. In this section we describe the data used, the calculation of the cepstral coefficients, model training and testing, and the results for single and simulated multi-vehicle experiments.

5.1 Database

The approach was tested on the US Army Research Laboratory's Acoustic-seismic Classification Identification Data Set (ACIDS) database. This database consists of various numbers of recordings from nine different vehicles. The number of recordings per vehicle varied from 7-60. The vehicles were recorded in three sperate environments: arctic, desert, and normal. The entire database was divided into two sections each consisting of approximately half of the total recordings for all nine vehicles. One section was used to train an HMM for each vehicle type and the other section was used to test the recognition performance of these models.

5.2 Preprocessing

The signal is first divided up into vectors consisting of 160 samples, which at the 1025.621 Hz sampling rate, corresponds to approximately 0.15 seconds. A spectral estimate is obtained for each vector using the "window method" of Grenander and Rosenblatt [8, 6]. In this method, a windowed autocorrelation estimate is Fourier transformed to obtain a spectral estimate. Depending on the window used, the method can ensure consistent spectral estimates are obtained. In [5], it is shown that cepstral estimates obtained from consistent spectral estimates are themselves consistent. The biased autocorrelation estimate is obtained from the inverse fast Fourier transform (FFT) of the magnitude squared of the FFT of the vector. The autocorrelation estimate is windowed by a Parzen window of length $K/3$. The Parzen window is an example of a window that results in a consistent spectral estimate [8, Section 6.2]. The windowed autocorrelation sequence is FFT'ed to form an estimate of the spectrum. The cepstral coefficients are obtained from the real part of the inverse FFT of the log spectrum. The zeroth order cepstral coefficient is discarded to accomplish gain invariance while the next 30 cepstral coefficients constitute the feature vector that is modelled by the HMM.

5.3 HMM Implementation

For vehicle classification, the HMM is trained for a particular vehicle type using examples of audio emissions from that vehicle type. Training involves estimating the parameter of the HMM. This parameter consists of the means and covariances of the Gaussian pdf's and the state transition matrix. Once an HMM has been trained for each vehicle type, recognition of a test signal from an unknown vehicle may be performed.

All vehicles are modelled by HMM's with $M = 8$ states. The Gaussian pdf's are non-zero mean with diagonal covariance matrices. Training of the HMM's is accomplished using a binary splitting procedure, where the number of states is doubled at each stage, until the desired number of 8 states is reached. At each stage, the parameter of the HMM is estimated using a two step procedure. In the first step, estimation is accomplished using the Baum-Viterbi algorithm [14]. In he second step, estimation is accomplished using the Baum algorithm [14]. This latter step yields a higher likelihood than assignment using the Baum-Viterbi only and may also yield a consistent parameter estimate [15]. The estimate can not be consistent if only the most likely state is used [16, 17]. Once the two steps are completed, each mixture is split into two, each with means that are slight perturbations of their parent's, and the two-step procedure repeated.

5.4 Single Vehicle Classification

Single vehicle classification can be performed by the MAP rule. This involves calculating the pdf of the test signal for each vehicle model. The results, in the form of a confusion matrix, appear in Figure 3. The correct classification rate, defined as the number of test signals correctly classified divided by the total number of test signals, is over 96%, which clearly demonstrate the capacity of the HMM to discriminate between ground vehicle types.

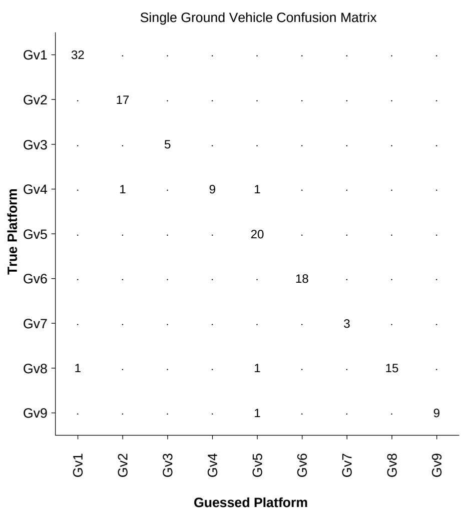

Figure 3: Ground Vehicle Confusion Matrix

5.5 Simulated Multiple Vehicle Classification

In real applications, ground vehicle classification systems must contend with more difficult situations than the single vehicle case represented in the experiment above. For example, sensors may record convoys consisting of an unknown number of vehicles of unknown types travelling in close proximity. A reasonable assumption in this situation is that signals from the vehicles are additive and independent, *i.e.* the emissions from one vehicle would not be affected by the emissions from any neighboring vehicles. In order to mimic these conditions, we synthesized recordings of multi-vehicle conveys by adding single vehicle emissions from the ACIDS database. For each recording in the testing section of the database, a recording from a different randomly chosen type of vehicle was added to the first recording. This combined signal was then presented to the same HMM classifier used previously.

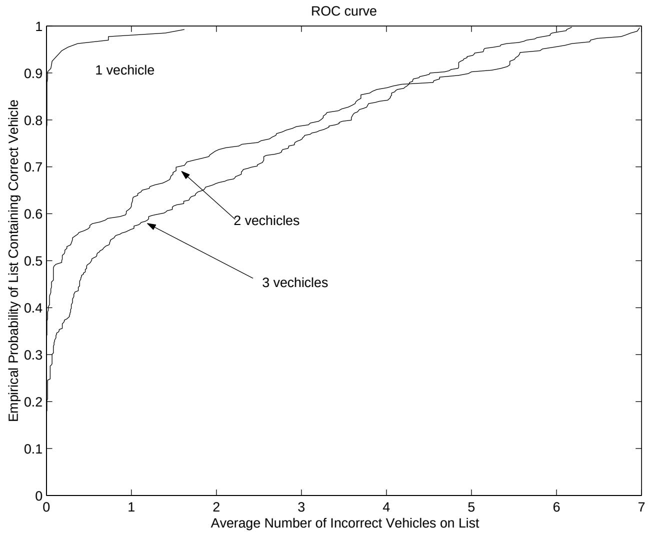

Figure 4: Ground Vehicle ROC curve for single and simulated multi-vehicle classification

For multi-vehicle classification, confusion matrices are an unwieldy way to present the results. For example, for two vehicles, the dimension of the confusion matrix is given by the number of possible ways to choose 2 vehicles from 10, *i.e.* 45. Here we present the results in the form of a multi-hypothesis receiver operating characteristic (ROC) curve discussed in section 13. Figure 4 shows the 2 vehicle ROC curve and for comparison, the single vehicle ROC curve is also shown. The curves demonstrate that the HMM classifier has the capacity to successfully identify two ground vehicles in the one recording, albeit with a significant decrease in performance compared to the single vehicle case. Reducing this performance gap is a primary aim of future work.

6 Comments

In the above experiment HMM's designed for single vehicles were used without modification in recognizing multiple vehicles. Improvements in performance would be expected using models representing multiple vehicles. These composite models may be obtained from single vehicle HMM's assuming the feature vectors are additive and statistically independent. Cepstral feature vectors do not have these properties. Feature vectors having these properties are those that model directly the time domain signal and those that model in the spectral domain. Future work will involve applying these representations to the problem.

As mentioned earlier, the signal, when multiple vehicles are present, is given by a superposition of individual vehicle signals. Estimating the number of individual signals in the recorded signal is an example of an order estimation problem. This same problem arises in various applications, for example, in estimating the number of harmonic components in a periodic signal or estimating the number of states in a hidden Markov model. Order estimation is a notoriously difficult estimation problem. Several well-established techniques applicable to certain situations are known. The general approach is to estimate the order that maximizes a penalized likelihood function of the recorded signal. The penalized likelihood function comprises the sum of the signal likelihood function for a given hypothesized order, and an additive penalty term for that order. The penalty term prevents overestimation of the order. We intend to implement and test vehicle counting via order estimation using various penalty terms including, but not necessarily limited to, the Bayesian information criterion (BIC) penalty term and the Akaike information criterion (AIC) penalty term.

A Empirical ROC curves

The x and y axes of the ROC curve are given by the values $\bar{N}$ and P_D respectively. In this appendix we show how to calculate these quantities empirically from test data. The dependency of these quantities on the threshold η is made explicit. Re-writing the expression for $\bar{N}$ we have

$$\bar{N}(\eta) = \sum_{j=1}^{J} \sum_{\substack{j'=1 \\ j' \neq j}}^{J} p(H_{j'}) \int \chi(\Lambda_j(y) > \eta) p(y|H_{j'}) dy \tag{14}$$

where

$$\chi(\Lambda_j(y) > \eta) = \begin{cases} 1 \text{ if } \Lambda_j(y) > \eta \\ 0 \text{ otherwise} \end{cases} \tag{15}$$

The quantities $p(y|H_{j'})$ and $p(H_{j'})$ are approximated by their empirical values from the testing set

$$\begin{aligned} p(y|H_{j'}) &\approx \frac{1}{N_{j'}} \sum_{n=1}^{N_{j'}} \delta(y - q_n^{(j')}) \\ p(H_{j'}) &\approx \frac{N_{j'}}{\sum_{j=1}^{J} N_j} \end{aligned} \tag{16}$$

where $q_n^{(j')}$ is the nth out of $N_{j'}$ testing signals from the j' hypothesis and $\delta(\cdot)$ is the Kronecker delta function. Substituting these values yields

$$\bar{N}(\eta) \approx \frac{1}{\sum_{j=0}^{J} N_j} \sum_{j=1}^{J} \sum_{\substack{j'=1 \\ j' \neq j}}^{J} \sum_{n=1}^{N_{j'}} \chi(\Lambda_j(q_n^{(j')}) > \eta) \tag{17}$$

Thus the empirical value of $\bar{N}$ for a given threshold is obtained by counting the number of times each signal in the test set appears in an incorrect decision region, and dividing by the total number of signals. Proceeding in a similar manner for $P_D(\eta)$ yields

$$P_D(\eta) \approx \frac{1}{\sum_{j=1}^{J} N_j} \sum_{j=1}^{J} \sum_{n=1}^{N_j} \chi(\Lambda_j(q_n^{(j)}) > \eta) \tag{18}$$

i.e. for a given threshold, count the number of signals that lie in their correct decision region, and divide by the total number of signals.

References

[1] G. D. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. on Information Theory*, vol. 14, no. 2, pp. 2062–20, Mar. 1968.

[2] Y. Ephraim, "Statistical model based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[3] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 39, pp. 2157–2166, Oct. 1991.

[4] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[5] N. Merhav and C. H. Lee, "On the asymptotic statistical behavior of empirical cepstral coefficients," *IEEE Trans. on Speech and Audio Processing*, vol. 41, no. 5, pp. 1990–1993, May 1993.

[6] Y. Ephraim and M. Rahim, "On second-order statistics and linear estimation of cepstral coefficients," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 2, pp. 162–176, Mar. 1999.

[7] J.-C. Junqua and J.-P. Haton, *Robustness in automatic speech recognition*, Kluwer Academic Publishers, Norwell, M.A., 1996.

[8] M. B. Priestely, *Spectral Analysis and Time Series*, Academic Press, 1992.

[9] J. D. Hamilton, *Time Series Analysis*, Princeton, 1994.

[10] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statistics*, vol. 37, pp. 1554–1563, Dec. 1966.

[11] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. Royal Stat. Soc.*, vol. B39, pp. 1–38, 1977.

[13] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, vol. I, Wiley, New York, 1968.

[14] Y. Ephraim, "Hidden Markov models," *The Encyclopedia of Operations Research*, 2002.

[15] B. G. Leroux, "Maximum likelihood estimation for hidden Markov models," *Stochastic Processes and Their Applications*, vol. 40, pp. pp 127–143, 1992.

[16] D. M. Titterington, "Comments on 'Application of conditional population-mixture model to image segmentation'," vol. 6, no. 5, pp. 656–657, Sept. 1984.

[17] P. Bryant and J. A. Williamson, "Asymptotic behaviour of classification maximum likelihood estimates," *Biometrika*, vol. 65, no. 2, pp. 273–281, 1978.